In-home detection of distress calls: the case of aged users

Frédéric Aman, Michel Vacher, Solange Rossato and François Portet

LIG, UMR5217 UJF/CNRS/Grenoble-INP/UMPF, 38041 Grenoble, France

{frederic.aman, michel.vacher, solange.rossato, francois.portet}@imag.fr

Abstract

In the context of technologies development aiming at helping aged people to live independently at home, the CIRDO[1] project aims at implementing an ASR system into a social inclusion product designed for elderly people in order to detect distress situations and provide capability to call for help. In this context we present a system able to detect distress and call for help sentences on line.

Index Terms: speech recognition, dependence, elderly, keyword detection

1. Introduction

A survey shows that 80% of people above 65 years old would prefer to stay living at home if they lose autonomy [1] and some of them needs the assistance of someone for regular elementary activities (nursing home services, domestic help, etc.) [2].

Few projects have seriously considered audio/speech technology in their design [3][4]. Part of this can be attributed to the complexity of setting up this technology in a real environment and to important challenges that still need to be overcome [5]. However, audio technology is well accepted by the potential users [6].

In this context, the CIRDO project promotes autonomy and support for elderly people by caregivers through the social inclusion product e-lio[2]. The objective of the project is to integrate an ASR system into this product that will include detection of distress situations and voice commands.

The use of elderly voice can be an issue for performance of speech recognition system. Some authors [7, 8] have reported that classical ASR systems exhibit poor performances with elderly voice because most acoustic models of ASR systems are acquired from non-aged voice samples. Indeed, ageing voice is characterized by some specific features such as imprecise production of consonants, tremors and slower articulation [9] due to changes in vocal production system [10, 11]. Howerver, it is possible to improve the recognition performances thanks to acoustic models adaptation [12].

This paper presents the ASR ability to detect sentences of distress in the case of ageing voice, the corpora and the ASR that we used were presented in [13].

2. Audio processing in line: CirdoX

We developed an application for the recognition of distress situations by analyzing the audio signal to detect sound and voice in real-time in an equipped flat. The system CirdoX will be coupled with an analysis of video scenes developed by the LIRIS laboratory in Lyon. Figure 1 shows the chain of acquisition and processing of the audio signal.

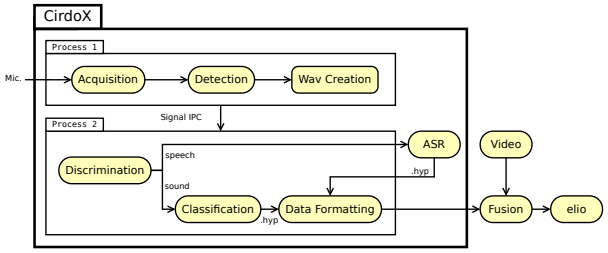

Figure 1: *Diagram of CirdoX.*

CirdoX is made of two parallel processes communicating by IPC (Inter-Process Communication) signals. This application is designed to be modular, with independent modules. For each module, the user can choose from among several plug-ins corresponding to different techniques.

The audio stream is captured by a microphone and recorded by the Acquisition module. This module can work through the plug-ins using the PortAudio library (audio card), Kinect or a National Instrument card. Then, the Detection module detects the occurrence of a signal (sound or voice). The detection of the occurrence of an audio event is based on the change of energy level of the three highest frequency coefficients of the Discrete Wavelet Transform (DTW) in a floating window frame (last 2048 samples without overlapping). Each time the energy on a channel goes beyond a self-adaptive threshold, an audio event is detected until the energy decrease below this level for at least an imposed duration.

Detection is done by thresholding on the discrete wavelet transform of the signal energy [14]. Then, a wave file is created by the Wave Creation module. The Discrimination module determines whether the signal is sound or speech. This module can be executed through the plug-ins using a Gaussian Mixture Model (GMM) or decision tree. In the case of sound, the Classification module defines to which sound class the signal belongs. It can work through GMM, decision tree and also Hidden Markov Model (HMM). In the case of speech, the signal is sent to the ASR module in order to output a sentence. Both hypotheses from Classification and ASR module are received by the Data Formatting module that sends a socket containing all the relevant data to the Fusion module. This one also receives data from the recognition of distress by video analysis in order to make a fusion between video and audio data. The Fusion module determines whether or not a distress situation is occurring and sends a socket containing its decision to the social inclusion product, e-lio. Then, e-lio executes an appropriate action, such as calling a doctor.

[1]This study was funded by the National Agency for Research under the project CIRDO - Industrial Research (ANR-2010-TECS-012)

[2]http://www.technosens.fr

3. Distress sentences detection

In this section, we focus now on a part of the Fusion module: the filtering of distress sentences. The filter detects the distress sentences into the ASR hypotheses by using Levenshtein distance [15]. Based on the GRePS laboratory's work [16] that interviewed elderly people in nursing homes to identify and describe what situations of distress they could have experienced, we created a list of sentences that can occur during a distress situation.

The filter calculates the Levenshtein distance between the ASR output hypothesis and all the distress sentences of the list. The sentence from the list with the best Levenshtein distance is selected according to a threshold. The lower the distance (score from 0 to infinity), the better the matching will be. To not be biased by orthography, the distance is calculated on a phonemic level. This approach takes into account some recognition errors such as word endings errors or slight variations. Moreover, in many cases, a miss-decoded word is phonetically close to the correct one due to close pronunciation.

A detected sentence occurs when the Levenshtein distance (normalized by the number of phonemes) is under the threshold, and is well matched if the ASR output hypothesis corresponds to the sentence selected from the list. If a detected sentence is well matched, it is considered as true positive (TP), otherwise a detected sentence not well matched is considered as a false positive (FP). A casual (C) ASR output hypothesis is never correct because of the ASR language model specifically adapted to distress utterances, and will never be well matched with a sentence of the distress sentences list. But if its distance is under the threshold, it is yet detected and becomes a false positive; if above threshold, it is true negative (TN). Also, if a distress (D) ASR output hypothesis is not well matched with the selected sentence and is not detected, it is also true negative. Finally, if a distress ASR output hypothesis is well matched with selected sentence but not detected, it is false negative (FN).

In order to assess the Levenshtein distance filter, we realized a decoding with the elderly speakers from the AD corpus, including 2796 distress sentences (same sentences as in Section 2 and 3) and 3006 casual sentences, for a total duration of 2 hour 12 minutes. The casual utterances were used as disrupters, with some sentences far from the distress ones: for instance *Les patates sont cuites* (Potatoes are cooked), or closer, for instance *Le médecin a appelé* (The doctor called). The average WER with the adapted acoustic model was 14.5% for the distress sentences, and was much higher for the casual sentences, 87.5%, due to the adapted-to-distress language model.

Then, we drew a ROC curve (Figure 2) representing the True Positive Rate (TPR, sensibility or recall) in function on the False Positive Rate (FPR, 1-specificity) by variating the threshold on the Levenshtein distance, for the 43 elderly speakers. At the point of equal error (cutoff=0.5), sensibility and specificity were equal to 91.8%. For our application, a high sensibility must be privileged in order to not miss some real distress situations. We obtained the positive and negative test showed in table 1 with threshold=0.5.

Table 1: *Positive and negative test.*

Threshold = 0.5	Right selection		False selection	
Distance $\leq$ threshold	TP = 2286		FP = 270	
	D = 2286	C = 0	D = 84	C = 186
Distance $>$ threshold	FN = 202		VN = 3044	
	D = 202	C = 0	D = 224	C = 2820

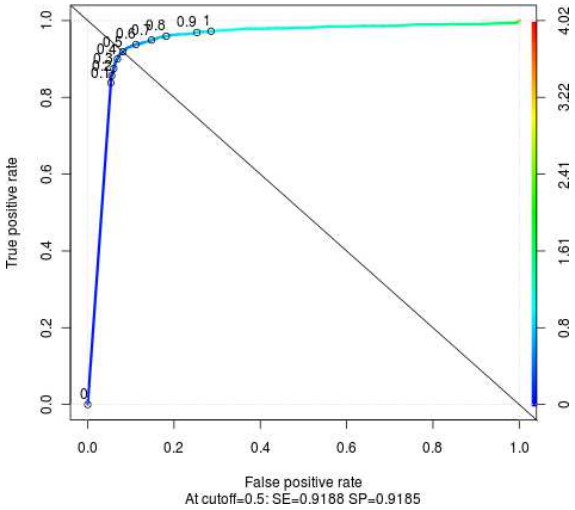

Figure 2: *ROC curve representing TPR in function of FPR.*

Also, we found a recall, precision and F-mesure equal to 91.8, 89.4 and 90.6%. This is in line with an other study conducted by Vacher et al.[17], where they detected domotic orders among casual utterances with a multisource ASR in a smart home environment. On their baseline system (best SNR channel), they found WER, recall, precision, and F-mesure respectively equal to 18.3, 88.0, 90.5 and 89.2%.

In table 2, we evaluated the distress sentences detection for the different levels[3] of elderly dependence GIR 2-3; GIR 4-5 and GIR 6 by using recall, precision and F-mesure. We observed that the system performance is quite depending on the level of the elderly dependence. Due to the lower F-mesure found on GIR 2-3 comparing to GIR 4-5 and GIR 6 (-7%), we suggest that such a system could be provided to elderly and usable by them only for score ranges from GIR 4 to GIR 6.

4. Conclusion

In this paper we presented the CirdoX software, used to filter and detect the distress calls. We evaluated a filter based on Levenshtein distance and showed that the recall and precision reach 91.8 and 89.4%. Moreover, we showed that such a system can not detect distress sentence well enough in the case of elderly in GIR 2-3 because of the high WER in this group.

In a future work, the whole system CirdoX including sound classification, speech recognition and fusion with video scene recognition is going to be evaluated in realistic condition in a real smart home [18]. Professional actors will play some scenarios, including for example falls to the ground because of the foot grasped in the carpet, sudden weakness, etc. We will assess how video can improve distress detection.

[3]We used a French national test as reference: the AGGIR (Autonomie Gérontologie Groupes Iso-Ressources) grid, http://vosdroits.service-public.fr/F1229.xhtml

Table 2: *Distress sentences detection in function of dependence*

	Recall	Precision	F-mesure
GIR 6	92.6%	91.1%	91.9%
GIR 4-5	92.0%	90.0%	91.0%
GIR 2-3	88.4%	80.7%	84.3%

5. References

[1] CSA, "Les français et la dépendance," http://www.csa.eu/fr/s26/nos-sondages-publies.aspx, 2003, accessed: 12/03/2013.

[2] C. Tlili, "Perspectives démographique et financières de la dépendance," *Rapport du groupe de travail sur la prise en charge de la dépendance*, 2011.

[3] M. Hamill, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems," *Journal of NeuroEngineering and Rehabilitation*, vol. 6, 2009.

[4] R. López-Cózar and Z. Callejas, "Multimodal dialogue for ambient intelligence and smart environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds. Springer US, 2010, pp. 559–579.

[5] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Development of audio sensing technology for ambient assisted living: Applications and challenges," *International Journal of E-Health and Medical Communications*, vol. 2, no. 1, pp. 35 – 54, march 2011.

[6] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[7] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Acoustic models of the elderly for large-vocabulary continuous speech recognition," *Electronics and Communications in Japan, Part 2*, vol. 87, pp. 49–57, 2004.

[8] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *9th International Conference on Speech Science and Speech Technology (InterSpeech 2008)*, Brisbane, Australia, 2008, pp. 2550–2553.

[9] W. Ryan and K. Burk, "Perceptual and acoustic correlates in the speech of males," *Journal of Communication Disorders*, vol. 7, pp. 181–192, 1974.

[10] N. Takeda, G. Thomas, and C. Ludlow, "Aging effects on motor units in the human thyroarytenoid muscle," *Laryngoscope*, vol. 110, pp. 1018–1025, 2000.

[11] P. Mueller, R. Sweeney, and L. Baribeau, "Acoustic and morphologic study of the senescent voice," *Ear, Nose, and Throat Journal*, vol. 63, pp. 71–75, 1984.

[12] M. Vacher, F. Portet, S. Rossato, F. Aman, C. Golanski, and R. Dugheanu, "Speech-based interaction in an aal context," *Gerontechnology*, vol. 11, p. 310, July 2012. [Online]. Available: http://dx.doi.org/10.4017/gt.2012.11.02.262.00

[13] F. Aman, M. Vacher, S. Rossato, and F. Portet, "Contribution à l'étude de la variabilité de la voix des personnes âgées en reconnaissance automatique de la parole (assessment of the acoustic models performance in the ageing voice case for asr system adaptation) [in french]," in *Actes de la conférence JEP-TALN-RECITAL 2012*, vol. 1: JEP, Grenoble, France, 2012, pp. 707–714.

[14] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat, "Information extraction from sound for medical telemonitoring," *Information Technology in Biomedicine, IEEE Transactions*, vol. 10, pp. 264–274, April 2006.

[15] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics-Doklady*, vol. 10, pp. 707–710, 1966.

[16] M.-E. B. Chaumon, B. Cuvillier, S. Bouakaz, and M. Vacher, "Démarche de développement de technologies ambiantes pour le maintien à domicile des personnes dépendantes : vers une triangulation des méthodes et des approches," in *Actes du 1er Congrès Européen de Stimulation Cognitive*, Dijon, France, 23-25 May 2012, pp. 121–122.

[17] M. Vacher, B. Lecouteux, and F. Portet, "Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment," in *EUSIPCO (European Signal Processing Conference)*, Bucarest, Romania, August 27-31 2012, pp. 1663–1667.

[18] M. Gallissot, J. Caelen, F. Jambon, and B. Meillon, "Une plateforme usage pour l'intégration de l'informatique ambiante dans l'habitat : Domus," *Technique et Science Informatiques (TSI)*, vol. 32, p. à paraître, 2013.